IT550: Information Retrieval

Assignment 10: Summarization

Instructor: Prasenjit Majumder

Learning Outcome: At the end of this assignment you will learn using extractive summarization algorithms and also evaluating summaries using ROUGE score

1 Problem description

Text summarization is the technique for generating a concise and precise summary of voluminous texts while focusing on the sections that convey useful information, and without losing the overall meaning.
Automatic text summarization aims to transform lengthy documents into shortened versions, something which could be difficult and costly to undertake if done manually.
Machine learning algorithms can be trained to comprehend documents and identify the sections that convey important facts and information before producing the required summarized texts.

Broadly, there are two approaches to summarizing texts:

1. **Extractive Summarization**: In this approach content is extracted from the original data, but the extracted content is not modified in any way. Examples of extracted content include key-phrases that can be used to "tag" or index a text document, or key sentences (including headings) that collectively comprise an abstract, and representative images or video segments, as stated above.

2. **Abstractive Summarization**: This approach builds an internal semantic representation of the original content, and then uses this representation to create a summary that is closer to what a human might express. Abstraction may transform the extracted content by paraphrasing sections of the source document, to condense a text more strongly than extraction.

2 Implementation

2.1 Dataset

- Use the BBC business news dataset for this assignment
- BBC business news contains news and summary in text format.
- Download the dataset from the following:
  https://drive.google.com/open?id=1epJmROcAV65GIgflnmJpXHXwkC5BYuU6

2.2 Exercise

- Implement LexRank and TexRank summarization approaches for generating the summary for all the news articles
- For each approach using the following sentence counts: 10,15,20,25
- For each approach and sentence count get the Rouge 1 and Rouge 2 scores for each document
- Average the scores obtained for all the documents for each approach and sentence count

3 References

- https://rare-technologies.com/text-summarization-in-python-extractive-vs-abstractive-techniques-re
- https://github.com/miso-belica/sumy

4 Submission

- You have to submit your assignment in Jupyter notebook with proper comments and explanation of your approach.
- For each of your approach you have to show the Rouge 1 and Rouge 2 scores
- The submission deadline for this assignment in **20th April 2020 at 11 PM**